[This Page Intentionally Left Blank]

6.0 ACCURACY OF THE 3T3 AND NHK NRU TEST METHODS

This section discusses the accuracy of the 3T3 and NHK NRU test methods for predicting acute oral systemic toxicity. Accuracy, the agreement between a test method result and an accepted reference value, is a critical component of the ICCVAM evaluation of the validation status of a test method (ICCVAM 2003). Although the 3T3 and NHK NRU test methods are not suitable as replacements for acute systemic toxicity assays, the ability of these assays to correctly predict LD_{50} values are used to evaluate their accuracy. The rationale for evaluating the accuracy of LD_{50} predictions is that the animal savings produced by using these *in vitro* test methods to predict starting doses for acute systemic toxicity assays will be greatest when the starting dose is as close as possible to the "true" LD_{50} value (see **Section 10** for the evaluation of animal savings).

The ability of the 3T3 and NHK NRU test methods to correctly predict rodent acute oral systemic toxicity is based on the validity of the *in vitro – in vivo* regression model. It is the *in vivo – in vitro* regression that establishes the relationship between the 3T3 and NHK NRU IC_{50} values and the predicted LD_{50} values that are to be used to set the starting doses for the acute oral systemic toxicity assays in this study.

Upon review of these regressions, it became apparent that the regression model could be improved. This section discusses the evolution of these improvements. Initially, since the regressions generated by the three laboratories were not statistically different, the data were combined (using a geometric mean IC_{50} of the three individual laboratory geometric mean IC_{50} values) into a single regression for each test method (3T3 and NHK). These regressions, in millimole units, were then compared to the RC millimole regression that was created using rat and mouse oral LD_{50} values from RTECS® and IC_{50} values from *in vitro* cytotoxicity assays using multiple cell lines and cytotoxicity endpoints for 347 substances with known molecular weights (Halle 1998). Because the 3T3 and NHK NRU test method regressions were not statistically different from the RC regression, the RC regression was chosen to predict the LD_{50} values from the NRU generated IC_{50} values because it is based on a much larger database.

The next steps taken were to improve upon the RC millimole regression's ability to accurately predict LD_{50} values from IC_{50} values, and to make the approach relevant to the testing of mixtures and substances without a known molecular weight in rats, the preferred species for acute oral toxicity testing (EPA 2002b; OECD 2001a; OECD 2001d). To achieve this goal, three new regressions are presented.

The first regression -- a RC rat-only millimole regression -- utilizes only the 282 substances in the RC dataset of 347 substances that had a reported rat LD_{50} value. The next step was to transform this RC rat-only millimole regression to one based on a weight basis (mg/kg body weight for LD_{50} and μg/mL for IC_{50}) in order to make the regression more generally applicable to the testing of mixtures and substances without a known molecular weight.

Upon review of this rat-only weight regression, it became apparent that many of the substances with underpredicted toxicity had mechanisms of toxicity that could not be expected to be detected in the 3T3 and NHK cell lines. These mechanisms included neurotoxic and cardiotoxic mechanisms, interference with energy utilization, and agents that alkylate macromolecules. Therefore, the third improved regression presented is based on an RC dataset of 232 substances that have rat LD_{50} data and that excludes the 50 substances which are reported to induce toxicity via one of the above mentioned mechanisms of action.

The ability of the 3T3 and NHK NRU IC_{50} data to correctly predict rat acute oral LD_{50} values, based on using the RC millimole regression and two of the modified regressions (RC rat-only weight regression and RC rat-only weight regression excluding substances with specific mechanisms of toxicity), was evaluated by determining the extent to which the appropriate GHS acute oral toxicity category was identified for each reference substance. This approach permits an assessment of accuracy specific to each GHS hazard classification category. The results of these analyses are presented in **Section 6.3**. The discordant reference substances from the predictions of GHS acute oral toxicity category are presented in **Appendix L-2**.

The remainder of **Section 6** discusses physical, chemical, and biological characteristics of substances that may have an impact on the accuracy of the 3T3 and NHK NRU test methods.

6.1 Accuracy of the 3T3 and NHK NRU Test Methods for Predicting Acute Oral Systemic Toxicity

Rodent LD_{50} values are used as the reference values for assessing the ability of the 3T3 and NHK NRU test methods to accurately predict acute oral systemic toxicity. The accuracy of the two *in vitro* cytotoxicity test methods is assessed in two ways: (1) by the goodness of fit of the *in vitro* NRU IC_{50} data to the rodent LD_{50} data in linear regression analyses, and (2) by the concordance (i.e., extent of agreement) between the GHS acute oral toxicity categories (UN 2005) assigned based on rodent LD_{50} data and those predicted using *in vitro* NRU IC_{50} data.

6.1.1 Linear Regression Analyses for the Prediction of *In Vivo* Rodent LD_{50} Values from *In Vitro* NRU IC_{50} Values

As described in **Section 5.3.4**, linear regressions for each test method were calculated using log IC_{50} values (mM) versus the corresponding reference log LD_{50} values (mmol/kg) identified in **Table 4-2**. The slopes for all regressions were statistically significantly different from zero ($p < 0.0001$), which indicates a significant relationship between *in vitro* IC_{50} values and the corresponding rodent LD_{50} values.

Comparison of the individual laboratory regressions to one another with the goodness of fit F-test described in **Section 5.3.3** (under "Generation of Other Linear Regressions") indicated that the laboratory-specific regressions for either *in vitro* NRU cytotoxicity test method were not significantly different from one another (see **Section 7.0** for a more detailed discussion of the results of this analysis). Because the individual laboratory regressions were not significantly different, data were combined into a single regression for each test method using the geometric mean of the mean IC_{50} values determined by each laboratory for each substance (see the "Combined-laboratory" regressions in **Table 6-1** and **Figure 6-1**). The

combined-laboratory 3T3 regression yielded a better fit to the reference LD_{50} data (adjusted $R^2 = 0.524$) than the combined-laboratory NHK regression (adjusted $R^2 = 0.455$).

Table 6-1 Linear Regression Analyses of the 3T3 and NHK NRU and *In Vivo* Rodent LD_{50} Test Results[a]

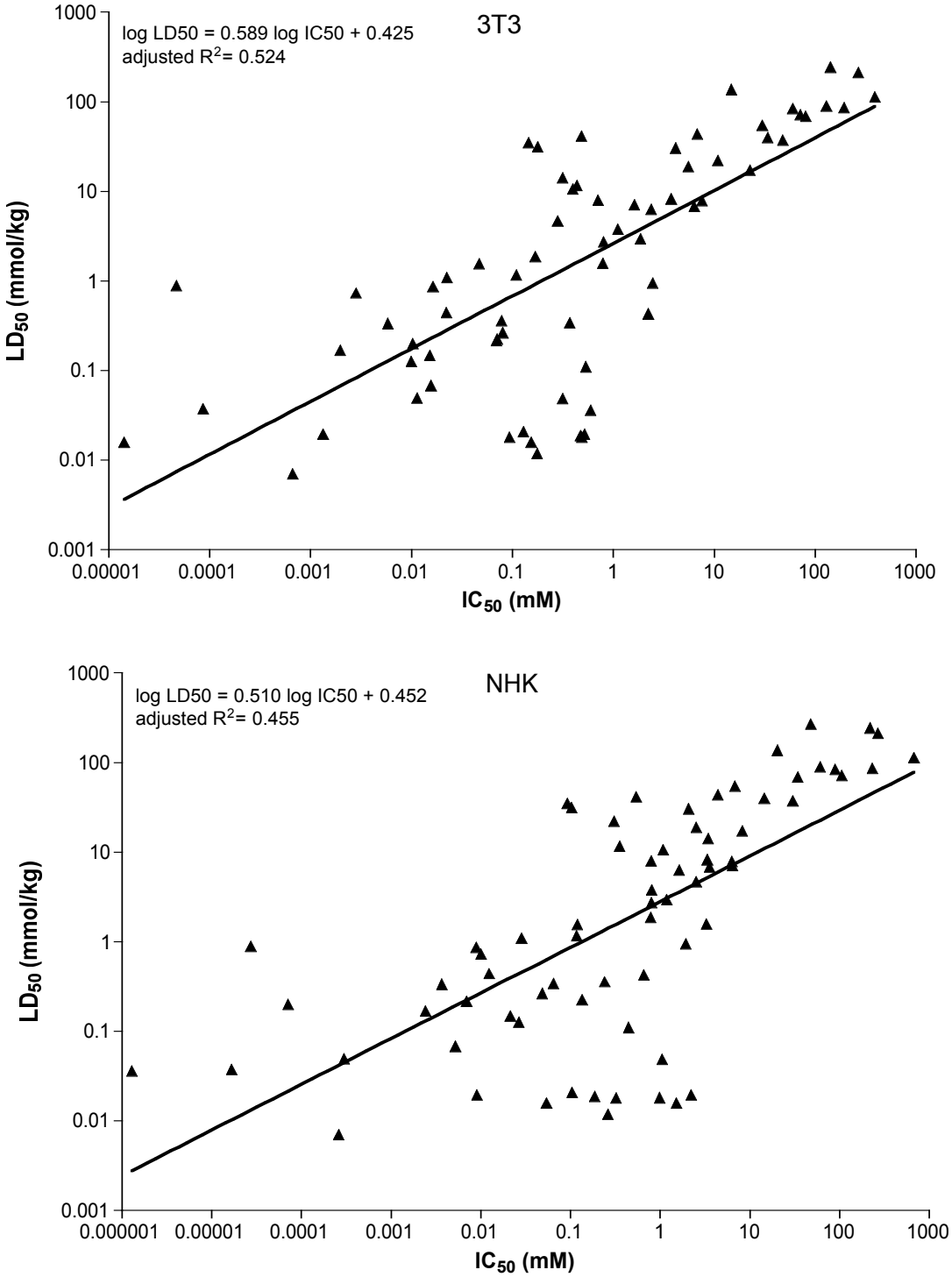
Laboratory	N[b]	Slope	Intercept	Adjusted R^2
3T3 NRU Test Method				
ECBC[c]	69	0.580	0.467	0.531
FAL[c]	67	0.543	0.287	0.432
IIVS[c]	69	0.585	0.467	0.534
Combined-laboratory[d]	70	0.589	0.425	0.524
NHK NRU Test Method				
ECBC[c]	69	0.507	0.405	0.446
FAL[c]	69	0.466	0.427	0.411
IIVS[c]	70	0.513	0.439	0.454
Combined-laboratory[d]	71	0.510	0.452	0.455

[a]Log IC_{50} in mM; log LD_{50} in mmol/kg.
[b]Number of substances used to calculate regression.
[c]Regression based on a single point per substance (i.e., the geometric mean of the within laboratory replicate IC_{50} values and the reference rodent oral LD_{50} from **Table 4-2**).
[d]Regression based on a single point per substance (i.e., the geometric mean of the geometric mean IC_{50} values obtained for each laboratory and the reference rodent oral LD_{50} from **Table 4-2**). Data for 70 substances in the 3T3 assay and 71 substances in the NHK assay. No laboratory achieved sufficient toxicity for calculation of an IC_{50} for carbon tetrachloride or methanol in the 3T3 NRU test method or for carbon tetrachloride in the NHK NRU test method.
Abbreviations: ECBC – US Army Edgewood Chemical Biological Center; FAL – FRAME Alternatives Laboratory; IIVS – Institute for *In Vitro* Sciences

6.1.2 Comparison of the Combined-Laboratory 3T3 and NHK NRU Regressions to the RC Millimole Regression

The NICEATM/ECVAM validation study tested 58 RC substances (see **Figure 3-1**). A comparison of the regression developed for the 3T3 and NHK NRU test results to the RC millimole regression was made to test the assumption of the *Guidance Document* that the RC millimole regression can be obtained with a basal cytotoxicity test method using a single cell type and cytotoxicity endpoint (ICCVAM 2001b). The regression for the 58 substances calculated using the RC IC_{50} and LD_{50} data points is shown in **Figure 6-2**. A graphical comparison of the RC millimole regression for the 58 substances to the 3T3 and NHK combined-laboratory regressions is shown in **Figure 6-3**. A statistical comparison of slope

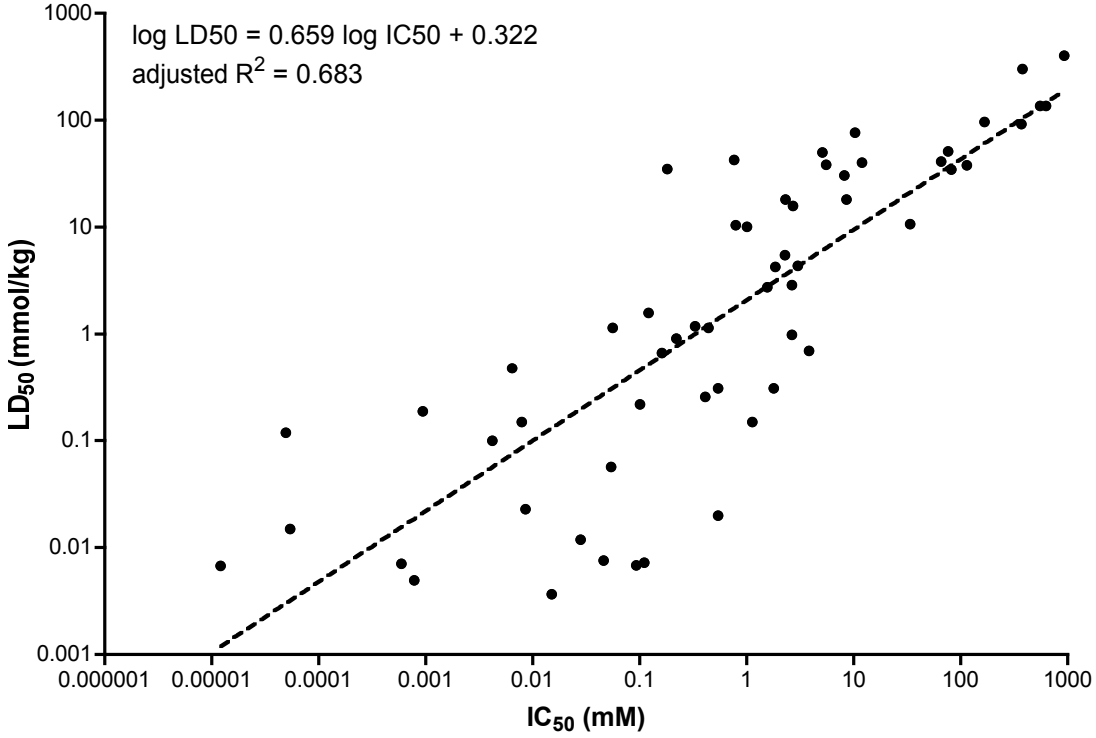
172 **Figure 6-1 Combined-Laboratory 3T3 and NHK Regressions**



173
174 Solid lines show the combined-laboratory regressions for each test method (see **Table 6-1**).
175

and intercept (simultaneously) using an F test showed that neither the 3T3 regression (p = 0.929) nor the NHK regression (p = 0.144) was different from the 58 RC substance regression.

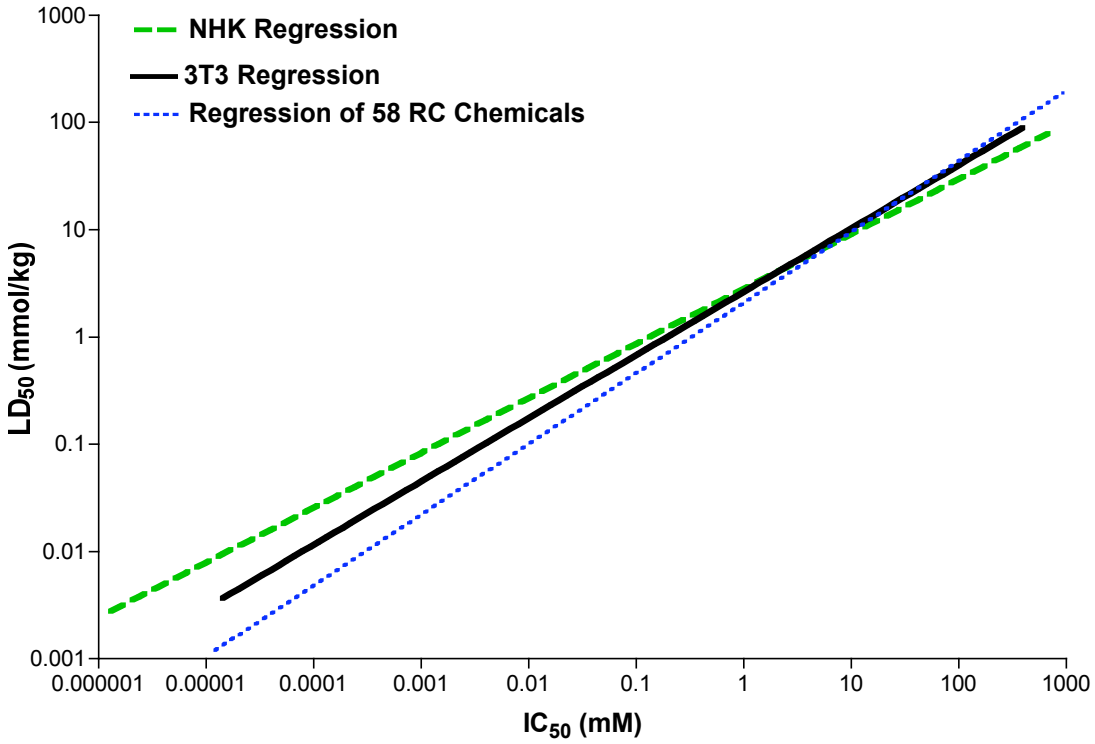
Figure 6-2 Regression for the 58 RC Substances Using RC Data



6.2 Improving the Prediction of *In Vivo* Rodent LD_{50} Values from *In Vitro* NRU IC_{50} Data

Since the RC and the 3T3 and NHK NRU IC_{50} – rodent acute oral LD_{50} regressions were not significantly different, the next step was an attempt to improve upon the RC millimole regression for the prediction of LD_{50} from IC_{50}. The RC data were used to develop three new regressions. For reference, the original RC millimole regression, log LD_{50} (mmol/kg) = 0.435 x log IC_{50x} (mM) + 0.625 (Halle 1998), is shown in **Table 6-2** and **Figure 6-4**.

Figure 6-3 Regression for the 58 RC Substances with the 3T3 and NHK NRU Regressions



Regression for the 58 RC substances using RC data is log LD_{50} = 0.659 log IC_{50} + 0.323 (adjusted R^2 = 0.683). The combined-laboratory 3T3 NRU regression, which uses data for 70 substances, is log LD_{50} = 0.589 log IC_{50} + 0.425 (adjusted R^2 = 0.524) (from **Table 6-1**). The combined-laboratory NHK NRU regression, which uses data for 71 substances, is log LD_{50} = 0.510 log IC_{50} + 0.452 (adjusted R^2 = 0.455) (from **Table 6-1**). No laboratory achieved sufficient toxicity for calculation of an IC_{50} for carbon tetrachloride or methanol in the 3T3 NRU test method or for carbon tetrachloride in the NHK NRU test method.

6.2.1 The RC Rat-Only Regression in Millimolar Units

The first regression used the RC data only for the 282 substances with rat LD_{50} data (i.e., the regression excluded the substances with mouse LD_{50} data) using the original units of mM for IC_{50} and mmol/kg for LD_{50} (see **Table 6-2** and **Figure 6-4**). Rat data only were used because:

- rats and mice may not have the same sensitivity to individual substances, regardless of the high correlation of a subset of 173 RC substances with both rat and mouse LD_{50} data (r_s = 0.88; $p < 0.0001$) (see **Section 4.1.4**)

- the majority of LD_{50} data used in the RC millimole regression were from studies using rats (282 rat data points and 65 mouse data points) (Halle 1998)
- the great majority of acute oral systemic toxicity testing is performed with rats

Table 6-2 Linear Regression Analyses to Improve the Prediction of Rodent LD_{50} from *In Vitro* NRU IC_{50} Using the RC Regression[a]

Data Used	Slope	Intercept	Adjusted R^2
347 RC substances with rat and mouse LD_{50} data – millimole units[c]	0.435	0.625	0.450[d]
282 RC substances with rat LD_{50} data – millimole units[c]	0.439	0.621	0.451
282 RC substances with rat LD_{50} data – weight units[e]	0.372	2.024	0.322
232 RC substances with rat LD_{50} data (excluded 50 substances with specific mechanisms of action[f]) – weight units[e]	0.357	2.194	0.353

[a]Slopes of all regressions were significantly different ($p < 0.05$) from zero at $p < 0.0001$.
[b]Simultaneous comparison of slopes and intercepts using an F test. Significance denoted by $p < 0.05$.
[c]IC_{50} in mM; LD_{50} in mmol/kg.
[d]Calculated from RC data (i.e., not reported by Halle [1998]).
[e]IC_{50} in μg/mL; LD_{50} in mg/kg.
[f]See the text for the applicable mechanisms and **Appendix K-3** for the identified substances.

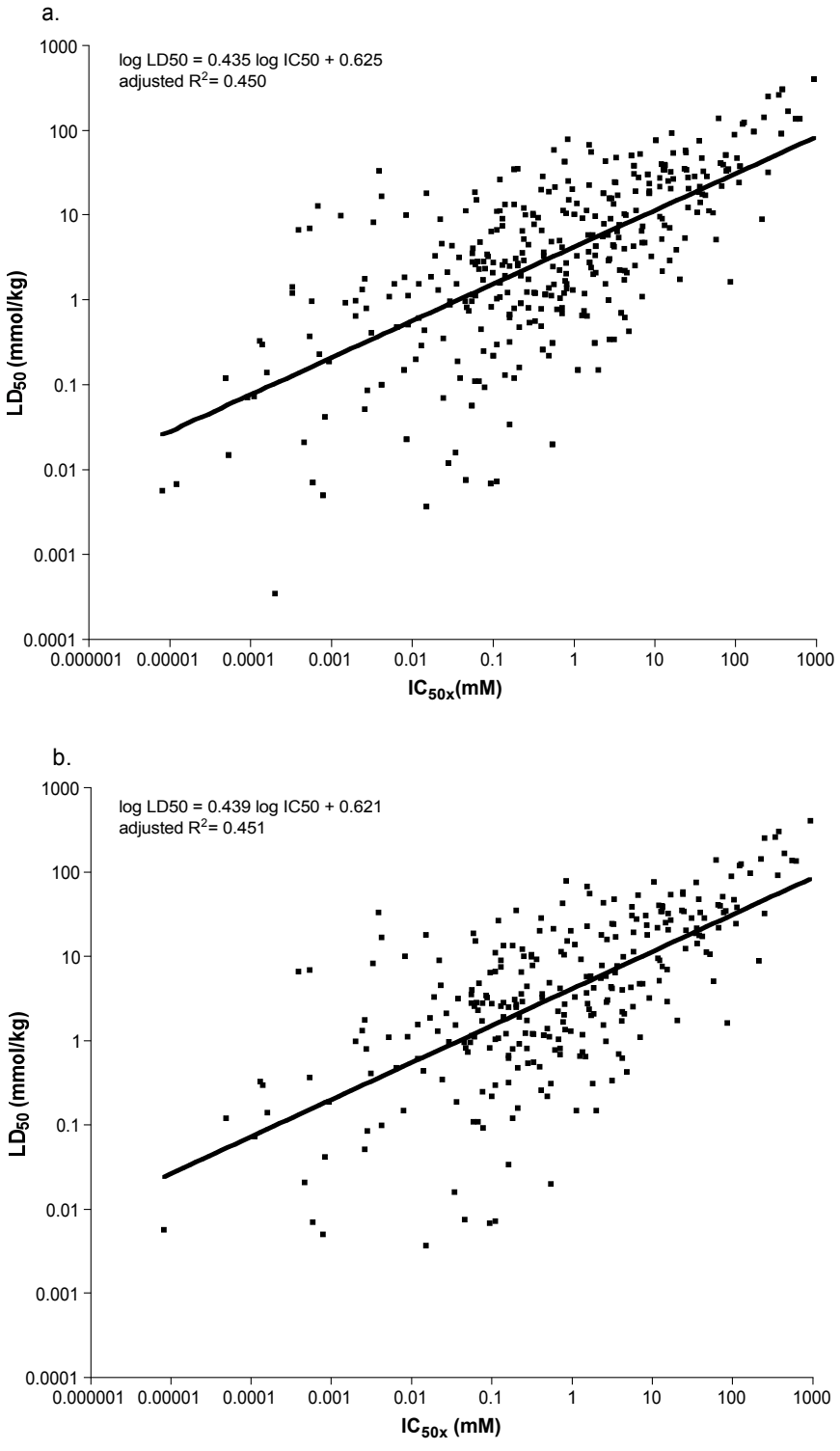
Table 6-2 shows that the regression using rat LD_{50} data only was almost identical to the original RC millimole regression which used both rat and mouse LD_{50} data. The slope changed from 0.435 for the RC millimole regression to 0.439 and the intercept changed from 0.625 to 0.621.

6.2.2 The RC Rat-Only Regression in Weight Units

The second regression used the same RC data for the 282 substances with rat LD_{50} data, but was calculated with weight units rather than millimolar units (see **Table 6-2** and **Figure 6-5**). Weight units (i.e., mg/kg for the LD_{50} and μg/mL for the IC_{50}) were selected for the units of measurement because

- millimole units are not applicable to mixtures and unknown substances
- they are most practical [i.e., in all regulatory systems, hazard classification is based on LD_{50} values expressed in mg/kg (see **Table 1-2**)]

235 **Figure 6-4 RC Regression (a) and RC Rat Regression (b) Using Millimole Units**



236

6.2.3 The RC Rat-Only Regression in Weight Units Excluding Substances with Specific Mechanisms of Toxicity

The third regression was a further refinement on the weight-unit regression developed from the 282 RC substances with rat LD_{50} data. It excluded the RC substances for which the mechanisms of toxic action were not expected to be active in the 3T3 and NHK cell cultures. This reduced the number of data points from 282 to 232 RC substances for the calculation of the regression (see **Table 6-2** and **Figure 6-5**). The third regression was significantly different ($p < 0.05$) from the RC rat-only weight regression when slopes and intercepts were simultaneously compared (F test; $p = 0.0063$). The idea for the further refinement for the rat RC millimole regression came from the evaluation of discordant substances (i.e., those greater than 0.699 or 0.5 log from the regression) when the 3T3 and NHK NRU data were used with the RC millimole regression (see **Appendix L-1**). For the 3T3 NRU, 13/30 (43%) of the discordant substances had mechanisms of toxicity that were not expected to be active in the 3T3 cell cultures. For the NHK NRU, 13/31 (42%) of the discordant substances had mechanisms of toxicity that were not expected to be active in the NHK cell cultures.

Development of the RC Rat-Only Weight Regression Excluding Substances with Specific Mechanisms of Toxicity

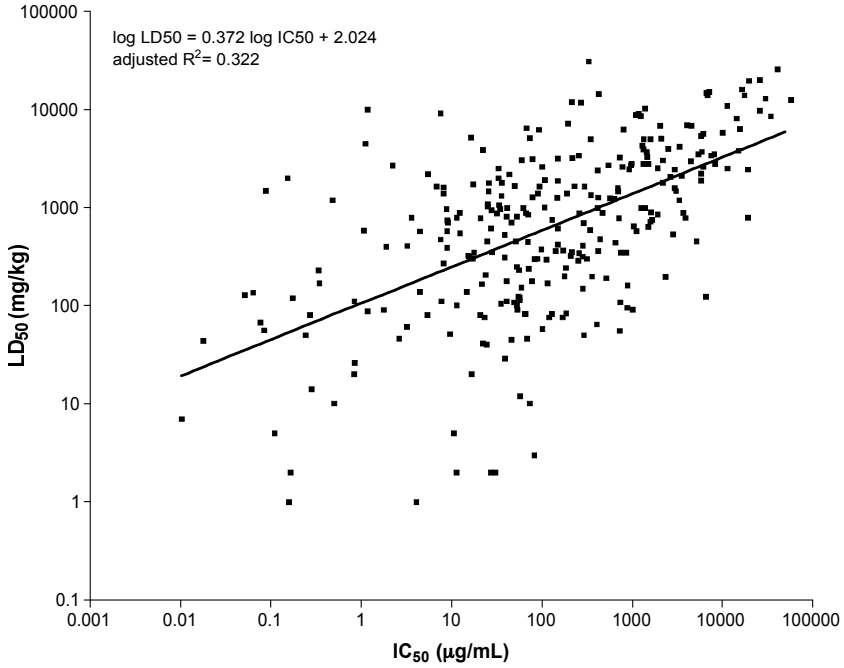
Mechanism of action data for the 282 RC substances with rat LD_{50} values were obtained from *Casarett & Doull's Toxicology* (Casarett et al. 2001) and the following Internet sources: HSDB (NLM 2005); Haz-Map (NLM 2005); Pesticide Action Network [PAN] Pesticides Database (PAN North America 2005); and IPCS INTOX Database (Canadian Centre for Occupational Health and Safety 2005) (see **Appendix K-3**). Mechanism of action information could not be found for all substances. For 35 of the 282 (12%) substances, only the product class could be identified; for seven (3%) substances, no information was found. Examination of the RC rat database revealed the following.

- Of the 282 substances, 73 (26%) were outliers[1] (i.e., log observed – log predicted $LD_{50} > 0.699$ as defined for the RC millimole regression).

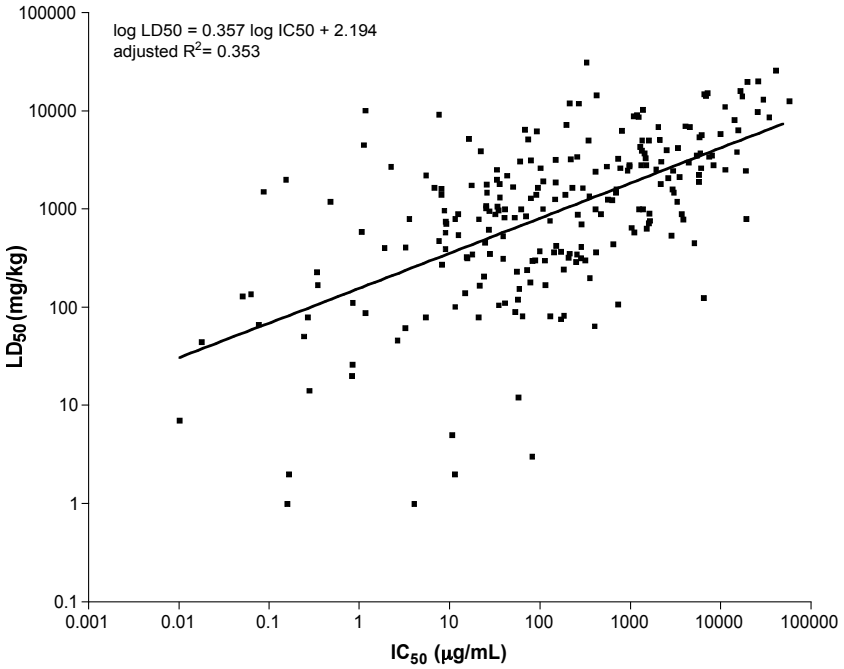
[1]Substance "outliers" are often referred to as discordant chemicals. Substance outliers are different from the replicate "outliers" described in **Sections 5.2** and **5.3**, which were extreme values in a set of replicate data. See **Section 13** for definitions.

Figure 6-5 RC Rat-Only Regression (a) and RC Rat-Only Regression after Excluding 50 Substances with Specific Mechanisms of Toxicity (b) Using Weight Units

a.



b.



- For 40 (55%) of the 73 substances, *in vivo* toxicity was underpredicted; for 33 (45%) of the 73 substances, *in vivo* toxicity was overpredicted.
- All underpredicted substances were very toxic, with $LD_{50} \leq 200$ mg/kg.
- The discordant status of 65% (26/40) of the underpredicted substances could be explained by four general mechanisms
    - neurotoxic (i.e., cholinesterase inhibitor, affects CNS nicotinic receptor, or otherwise neurotoxic by a mechanism other than membrane destabilization such as that produced by a solvent)
    - interferes with energy utilization (i.e., interferes with ATP synthesis, inhibits ADP phosphorylation, or uncouples oxidative phosphorylation, or is a metabolic poison)
    - cardiotoxic via specific mechanisms (i.e., positive inotropic action, calcium channel blocker)
    - alkylates cellular proteins and other macromolecules (i.e., covalently binds to enzymes and other proteins to disrupt normal function)

Substances with such mechanisms would not be expected to exert their toxic mechanisms in the 3T3 and NHK cells and thus, they would be expected to fit the RC millimole regression poorly, as evidenced by their discordant status. A new regression was calculated after the exclusion of all substances in the RC database known to act by these four mechanisms; this included the 26 underpredicted substances and 24 other substances that were not identified as outliers. The substances excluded from the RC rat weight regression are identified in **Appendix K-3**.

6.3 Accuracy of the 3T3 and NHK NRU Test Methods for Toxicity Category Predictions

The 3T3 and NHK NRU test methods are not suitable as replacements for acute oral systemic toxicity assays. However, the use of *in vitro* NRU test methods to reduce animal use for acute oral systemic toxicity assays depends upon their accuracy for the prediction of LD_{50} values. NRU-predicted LD_{50} values were determined by using the *in vitro* NRU IC_{50} values

in the regressions presented in **Table 6-2.** The predicted LD_{50} values were then used to assign each substance to a predicted GHS acute oral category (UN 2005). The accuracy of the 3T3 and NHK NRU test methods for predicting GHS toxicity categories was determined by comparison with categorization based on *in vivo* rodent LD_{50} data. This accuracy evaluation approach was used because the regulatory use of acute systemic toxicity test results is for the purpose of hazard classification and labelling of products to protect handlers and consumers.

The following regressions from **Section 6.2** were evaluated for accuracy of GHS oral toxicity category predictions:

- RC millimole regression
- RC rat-only weight regression
- RC rat-only weight regression excluding substances with specific mechanisms of toxicity

The regression calculated using the rat RC data in millimole units (**Section 6.2.1**) was not evaluated separately since it was very similar to the original RC millimole regression, which used both rat and mouse data. As **Table 6-2** shows, the slopes and intercepts varied only in the thousandths digits.

Data for the same reference substances were evaluated for each regression. Forty-six substances were evaluated for the 3T3 NRU test method and 47 substances were evaluated for the NHK NRU test method. Of the original 72 substances tested, epinephrine bitartrate, colchicine, and propylparaben were excluded because they were removed from the calculation of the RC rat-only weight regression due to the lack of rat oral reference LD_{50} data. The 21 substances with specific mechanisms of toxicity in **Table 6-3** were excluded from all analyses to be consistent with those removed from the RC rat-only weight regression excluding substances with specific mechanisms of toxicity. These substances have known mechanisms of toxicity that are not expected to be active in the 3T3 and NHK cell cultures.

Table 6-3 Substances Deleted from the Evaluations of the 3T3 and NHK NRU Test Methods and Regressions Due to Mechanisms of Toxicity Not Expected to Be Active in the 3T3 and NHK Cell Cultures

Substance	Mechanism of Toxicity[1]
Neurotoxic	
Amitriptyline HCl	Blocks norepinephrine, 5-hydroxytryptamine, and dopamine presynaptic uptake; prevents reuptake of heart norepinephrine.
Atropine sulfate	Antimuscarinic, anticholinergic action; competitive antagonism of anticholinesterase at cardiac & CNS receptor sites.
Caffeine	Inhibition of phosphodiesterase leading to AMP accumulation, translocation of intracellular Ca^{++}, adenosine receptor antagonism, neurotoxic.
Carbamazepine	Therapeutically decreases firing of noradrenergic neurons.
Chloral hydrate	Potentiation of $GABA_A$ receptor activity, inhibition of N-methyl-D-aspartate activity, & modulation of 5-hydroxytryptamine$_3$ receptor-mediated depolarization of the vagas nerve[2].
Dichlorvos	Inhibition of acetylcholinesterase resulting in acetylcholine accumulation in CNS & effector organs.
Disulfoton	Inhibition of acetylcholinesterase resulting in acetylcholine accumulation in CNS & effector organs.
Endosulfan	Affects brain neurotransmitter levels[3].
Fenpropathrin	Delays closure of sodium channel causing persistent depolarization of membrane.
Glutethimide	CNS depression, anticholinergic activity.
Haloperidol	Blocks dopamine receptors.
Lindane	CNS depression through inhibition of GABA receptor linked chloride channel at the picrotoxin binding site, leading to blockade of chloride influx into neurons.
Nicotine	Cholinergic block causing polarization of CNS and PNS synapses.
Parathion	Inhibition of acetylcholinesterase resulting in acetylcholine accumulation in CNS & effector organs.
Phenobarbital	CNS depression through inhibition of GABA synapses, inhibits hepatic NADH cytochrome oxidoreductase.
Physostigmine	Inhibition of acetylcholinesterase resulting in acetylcholine accumulation in CNS & effector organs.
Strychnine	Increases glutamic acid in the CNS.
Interferes with Energy Utilization	
Paraquat	Multisystem failure due to depletion of superoxide dismutase, with formation of free radicals & lipid peroxidation; lung fibrosis due to accumulation.
Potassium cyanide	General enzyme inhibition, high affinity for Fe^{+++}, inhibits cell respiration by inhibition of cytochrome oxidase.
Cardiotoxic	
Procainamide HCl	Slows impulse conduction in the heart[4].
Verapamil HCl	Inhibition of transmembrane Ca^{++} flux in excitatory tissues, alpha-adrenergic blockade.

Abbreviations: NA = not available or information not found; CNS = central nervous system; GABA = gamma aminobutyric acid; PNS = peripheral nervous system; NADH = nicotine adenine dinucleotide (reduced).
[1]Ekwall et al. (1998) or Hazardous Substances Data Bank (NLM 2001, 2002) unless otherwise noted.
[2]EPA (2000b).
[3]ATSDR (2000a).
[4]Hardman et al. (1996).

Carbon tetrachloride and methanol were excluded from the 3T3 NRU evaluations because no laboratory attained sufficient toxicity in any test for the calculation of an IC_{50}. Carbon tetrachloride was also excluded from the NHK NRU evaluations because no laboratory attained sufficient toxicity in any test for the calculation of an IC_{50}.

The tables providing accuracy information in this section (**Tables 6-4** to **6-6**) are divided into top and bottom parts that provide accuracy data for the 3T3 and NHK NRU test methods, respectively. For each part, the toxicity categories corresponding to the *in vivo* LD_{50} data are provided in rows that are labeled on the far left side of the table. The toxicity categories predicted by the *in vitro* NRU assays (and associated regressions) are provided in columns that are labeled across the top of each part (i.e., 3T3 or NHK NRU-predicted toxicity category) of the table. The numbers at the intersections of the *in vivo* LD_{50} rows and 3T3 or NHK NRU-predicted toxicity category columns are the numbers of substances with *in vitro* category predictions that correspond to the various *in vivo* categories. The right sides of the tables also provide summaries containing, for each *in vivo* toxicity category and for the total number of substances evaluated: number of substances, the accuracy of the 3T3 or NHK NRU prediction, and the percentage of substances for which toxicity has been overpredicted and underpredicted by the *in vitro* NRU methods. In each of the 3T3 NRU and NHK NRU sections of the table, a summary of predictivity[2] is also provided for each predicted toxicity category along with the percentage of substances with category (i.e., toxicity) underpredicted and overpredicted.

6.3.1 Prediction of Toxicity Category by the 3T3 and NHK NRU Test Methods Using the RC Millimole Regression

Table 6-4 shows the concordance of the observed (i.e., *in vivo*) and predicted GHS acute oral toxicity categories (UN 2005) for each *in vitro* NRU cytotoxicity test method using the geometric mean IC_{50} values (of the three laboratories) in the RC millimole regression, log LD_{50} (mmol/kg) = 0.435 x log IC_{50} (mM) + 0.625. Accuracy is the agreement of the

[2] Proportion of *in vivo* category matches for all substances with *in vitro* predictions for a particular category. Predictivity is an indicator of test accuracy (ICCVAM 2003).

365 **Table 6-4 Prediction of GHS Toxicity Category[1] by the 3T3 and NHK NRU Test Methods and the**
366 **RC Millimole Regression[2]**

Reference Rodent LD_{50}[3]	3T3 NRU-Predicted Toxicity Category						Total	Accuracy	Toxicity Overpredicted	Toxicity Underpredicted
	< 5	5 – 50	50 – 300	300 – 2000	2000 – 5000	> 5000				
< 5	0	3	1	3	0	0	7[4]	0%	0%	100%
5 –50	0	1	3	1	1	0	6[5]	17%	0%	83%
50 – 300	0	0	4	2	0	0	6[6]	67%	0%	33%
300 – 2000	0	0	0	6	0	0	6[7]	100%	0%	0%
2000 – 5000	0	0	0	11	0	0	11[8]	0%	100%	0%
> 5000	0	0	0	6	3	1	10[9,10]	10%	90%	0%
Total	0	4	8	29	4	1	46	26%	43%	30%
Predictivity	0%	25%	50%	21%	0%	100%				
Category Underpredicted	0%	0%	0%	59%	75%	0%				
Category Overpredicted	0%	75%	50%	21%	25%	0%				
Reference Rodent LD_{50}[3]	**NHK NRU-Predicted Toxicity Category**						**Total**	**Accuracy**	**Toxicity Overpredicted**	**Toxicity Underpredicted**
	< 5	**5 – 50**	**50 – 300**	**300 – 2000**	**2000 – 5000**	**> 5000**				
< 5	0	1	3	2	1	0	7[4]	0%	0%	100%
5 – 50	0	3	3	0	0	0	6[5]	50%	0%	50%
50 – 300	0	1	3	2	0	0	6[6]	50%	17%	33%
300 – 2000	0	0	0	6	0	0	6[7]	100%	0%	0%
2000 – 5000	0	0	0	10	1	0	11[8]	9%	91%	0%
> 5000	0	0	0	6	5	0	11[10]	0%	100%	0%
Total	0	5	9	26	7	0	47	28%	47%	26%
Predictivity	0%	60%	33%	23%	14%	0%				
Category Underpredicted	0%	20%	0%	62%	71%	0%				
Category Overpredicted	0%	20%	67%	15%	14%	0%				

[1]GHS-Globally Harmonized System of Classification and Labelling of Chemicals with LD_{50} in mg/kg (UN 2005).
< 5: $LD_{50} \leq 5$ mg/kg
5 – 50: $5 < LD_{50} \leq 50$ mg/kg
50 – 300: $50 < LD_{50} \leq 300$ mg/kg
300 – 2000: $300 < LD_{50} \leq 2000$ mg/kg
2000 – 5000: $2000 < LD_{50} \leq 5000$ mg/kg
> 5000: $LD_{50} > 5000$ mg/kg
[2]The RC millimole regression is log LD_{50} (mmol/kg) = log IC_{50} (mM) X 0.435 + 0.625. Numbers in table represent number of substances.
[3]Reference oral LD_{50} values from **Table 3-2**.
[4]Epinephrine bitartrate excluded because no rat LD_{50} was identified. Disulfoton, parathion, strychnine and physostigmine excluded based on mechanism of toxicity (see **Table 6-3**).
[5]Colchine excluded because no rat LD_{50} was identified. Dichlorvos, endosulfan, fenpropathrin, nicotine, and potassium cyanide excluded based on mechanism of toxicity (see **Table 6-3**).
[6]Caffeine, haloperidol, lindane, paraquat, phenobarbital, and verapamil HCl excluded based on mechanism of toxicity (see **Table 6-3**).
[7]Amitriptyline, atropine sulfate, carbamazepine, chloral hydrate, glutethimide, and procainamide HCl excluded based on mechanism of toxicity (see **Table 6-3**).
[8]Carbon tetrachloride excluded because no laboratory attained sufficient toxicity for the calculation of an IC_{50}.
[9]Methanol excluded because no laboratory attained sufficient toxicity for the calculation of an IC_{50}.
[10]Propylparaben excluded because no rat LD_{50} was identified.

category predictions with those based on the reference rodent LD_{50} values in **Table 3-2**, which are the values used for the original classification of the test substances. Substances for which the *in vitro* toxicity category prediction does not match the *in vivo* determined toxicity category are considered discordant substances for the GHS toxicity category predictions.

In Vitro – In Vivo Concordance Using the RC Millimole Regression

The overall accuracy of the 3T3 NRU test method for correctly predicting GHS toxicity classification category using the RC millimole regression was 26% (12/46 substances) (**Table 6-4**). *In vivo* toxicity was overpredicted for 43% (20) and underpredicted for 30% (14) of the 46 substances. For this analysis, in terms of each GHS toxicity classification category:

- 0 (0%) of seven substances with $LD_{50} < 5$ mg/kg was correctly predicted
- 1 (17%) of six substances in the $5 < LD_{50} \leq 50$ mg/kg category was correctly predicted
- 4 (67%) of six substances in the $50 < LD_{50} \leq 300$ mg/kg category were correctly predicted
- 6 (100%) of six substances in the $300 < LD_{50} \leq 2000$ mg/kg category were correctly predicted; however, this toxicity category was also predicted for 23 other substances (79%; 23/29) that did not match this category *in vivo*. Thus, the predictivity for this category was 21% (6/29 substances predicted for this category matched the *in vivo* category).
- 0 (0%) of the 11 substances in the $2000 < LD_{50} \leq 5000$ mg/kg category were correctly predicted
- 1 (10%) of the 10 substances in the $LD_{50} > 5000$ mg/kg range was correctly predicted

The overall accuracy of the NHK NRU cytotoxicity test method for correctly predicting the GHS toxicity classification, when the prediction was based on the RC millimole regression, was 28% (13/47 substances) (see **Table 6-4**). Toxicity was overpredicted for 47% (22) and underpredicted for 26% (12) of the 47 substances. The pattern of concordance between *in vitro* and *in vivo* results for the NHK assay with the RC millimole regression was similar to

the 3T3 results with the exception that two more substances were correctly in the $5 < LD_{50} \leq 50$ mg/kg category. For this analysis, in terms of each GHS toxicity classification category:

- 0 (0%) of seven substances with $LD_{50} < 5$ mg/kg were correctly predicted
- 3 (50%) of six substances in the $5 < LD_{50} \leq 50$ mg/kg and in the $50 < LD_{50} \leq 300$ mg/kg categories were correctly predicted
- 6 (100%) of six substances in the $300 < LD_{50} \leq 2000$ mg/kg category were correctly predicted; however, this toxicity category was also predicted for 20 (77%; 20/26) other substances with *in vivo* data that did not match the category. Thus, the predictivity for this category was 23%.
- 1 (9%) of 11 substances in the $2000 < LD_{50} \leq 5000$ mg/kg category was correctly predicted
- 0 (0%) of 11 substances in the $LD_{50} > 5000$ mg/kg range were correctly predicted.

For both *in vitro* NRU cytotoxicity test methods, when predicted GHS toxicity categories did not match the reference rodent GHS toxicity category, the RC millimole regression generally underpredicted toxicity for substances in the highest toxicity (i.e., lowest LD_{50}) categories and overpredicted toxicity for substances in the lowest toxicity (i.e., highest LD_{50}) categories (see **Table 6-4**). While substances at the very low and very high ends of the toxicity range were poorly predicted, those in the middle of the toxicity range (i.e., $300 < LD_{50} \leq 2000$ mg/kg) were predicted quite well.

Discordant Substances for Prediction of Toxicity Category by the 3T3 and NHK NRU Test Methods and the RC Millimole Regression

Appendix L-2 identifies the discordant substances for which the *in vitro* predicted GHS toxicity category did not match the GHS toxicity category assigned based on the reference rodent LD_{50} data in **Table 3-2**. For the 3T3 NRU test method, the toxicity was underpredicted for 14 (30%) and overpredicted for 20 (43%) of the 34 discordant substances. For the NHK NRU test method, toxicity was underpredicted for 12 (35%) and overpredicted for 22 (65%) of the 34 discordant substances. The fact that there were more substances for which toxicity was overpredicted is a result of the removal of substances with specific

mechanisms of toxicity that were not expected to be active in the 3T3 and NHK cell cultures. The toxicity for most of these substances would have been underpredicted. **Figure 3-1** shows that most of the RC substances selected for testing in the NICEATM/ECVAM validation study are below the RC millimole regression line. Thus, the RC millimole regression is expected to predict lower toxicity (i.e., a higher LD_{50}) for these substances.

6.3.2 Prediction of Toxicity Category by the 3T3 and NHK NRU Test Methods Using the RC Rat-Only Weight Regression

Table 6-5 shows the concordance of the observed and predicted GHS acute oral toxicity categories for each *in vitro* NRU test method using the geometric mean IC_{50} values (of the three laboratories) and the RC rat only weight regression from **Table 6-2**. The regression formula for the RC rat-only weight regression was log LD_{50} (mg/kg) = log IC_{50} (μg/mL) x 0.372 + 2.024. Accuracy is the agreement of the *in vitro* NRU cytotoxicity GHS toxicity category predictions with those based on the reference rat oral LD_{50} values from **Table 4-2**.

In Vitro – In Vivo Concordance with the RC Rat-Only Weight Regression

The overall accuracy of the 3T3 NRU test method with the RC rat-only weight regression was 35% (16) for the results from 46 substances (**Table 6-5**). The toxicity was overpredicted for 41% (19) and underpredicted for 24% (11) of the 46 substances. For this analysis, in terms of each GHS toxicity classification category:

- 0 (0%) of four substances with $LD_{50} < 5$ mg/kg were correctly predicted
- 1 (14%) of seven substances in the $5 < LD_{50} \leq 50$ mg/kg GHS toxicity category was correctly predicted
- 4 (80%) of five substances in the $50 < LD_{50} \leq 300$ mg/kg GHS toxicity category were correctly predicted; however, since seven other substances were also predicted for this category, the predictivity was 36% (4/11)
- 7 (78%) of nine substances in the $300 < LD_{50} \leq 2000$ mg/kg GHS toxicity category were predicted correctly. Since a total of 22 substances were predicted for this category, the predictivity was 32% (7/22).

Table 6-5 Prediction of GHS Toxicity Category[1] by the RC Rat-Only Weight Regression[2]

Reference Rodent LD_{50}[3]	3T3 NRU-Predicted Toxicity Category						Total	Accuracy	Toxicity Overpredicted	Toxicity Underpredicted
	< 5	5 – 50	50 – 300	300-2000	2000-5000	> 5000				
< 5	0	0	2	2	0	0	4[4]	0%	0%	100%
5 – 50	0	1	4	2	0	0	7[5]	14%	0%	86%
50 – 300	0	0	4	1	0	0	5[6]	80%	0%	20%
300 – 2000	0	1	1	7	0	0	9[7]	78%	22%	0%
2000 – 5000	0	0	0	5	4	0	9[8]	44%	56%	0%
> 5000	0	0	0	5	7	0	12[9,10]	0%	100%	0%
Total	0	2	11	22	11	0	46	35%	41%	24%
Predictivity	0%	50%	36%	32%	36%	0%				
Category Underpredicted	0%	50%	9%	46%	64%	0%				
Category Overpredicted	0%	0%	55%	23%	0%	0%				
Reference Rodent LD_{50}[3]	**NHK NRU-Predicted Toxicity Category**						**Total**	**Accuracy**	**Toxicity Overpredicted**	**Toxicity Underpredicted**
	< 5	**5 – 50**	**50 – 300**	**300 – 2000**	**2000 – 5000**	**> 5000**				
< 5	0	1	2	1	0	0	4[4]	0%	0%	100%
5 – 50	0	1	4	2	0	0	7[5]	14%	0%	86%
50 – 300	0	1	3	1	0	0	5[6]	60%	20%	20%
300 – 2000	0	1	0	8	0	0	9[7]	89%	11%	0%
2000 – 5000	0	0	0	8	1	0	9[8]	11%	89%	0%
> 5000	0	0	0	6	6	1	13[10]	8%	92%	0%
Total	0	4	9	26	7	1	47	30%	47%	23%
Predictivity	0%	25%	33%	31%	14%	100%				
Category Underpredicted	0%	50%	0%	54%	86%	0%				
Category Overpredicted	0%	25%	67%	15%	0%	0%				

[1]Globally Harmonized System of Classification and Labelling of Chemicals with LD_{50} in mg/kg (UN 2005).

< 5: $LD_{50} \leq 5$ mg/kg

5 – 50: $5 < LD_{50} \leq 50$ mg/kg

50 – 300: $50 < LD_{50} \leq 300$ mg/kg

300 – 2000: $300 < LD_{50} \leq 2000$ mg/kg
2000 – 5000: $2000 < LD_{50} \leq 5000$ mg/kg
> 5000: $LD_{50} > 5000$ mg/kg

[2]The RC rat-only weight regression is log LD_{50} (mg/kg) = log IC_{50} (μg/mL) X 0.372 + 2.024.
[3]Reference rodent LD_{50} values from **Table 4-2**.
[4]Epinephrine bitartrate excluded because no rat LD_{50} was identified. Disulfoton and physostigmine excluded based on mechanism of toxicity (see **Table 6-3**).
[5]Colchine excluded because no rat LD_{50} was identified. Endosulfan, parathion, potassium cyanide, and strychnine excluded based on mechanism of toxicity (see **Table 6-3**).
[6]Dichlorvos, fenpropathrin, lindane, paraquat, phenobarbital, nicotine, and verapamil HCl excluded based on mechanism of toxicity (see **Table 6-3**).
[7]Amitriptyline, atropine sulfate, caffeine, chloral hydrate, glutethimide, haloperidol, and procainamide HCl excluded based on mechanism of toxicity (see **Table 6-3**).
[8]Carbon tetrachloride excluded because no laboratory attained sufficient toxicity for the calculation of an IC_{50}. Carbamazepine excluded based on mechanism of toxicity (see **Table 6-3**).
[9]Methanol excluded because no laboratory attained sufficient toxicity for the calculation of an IC_{50}.
[10]Propylparaben excluded because no rat LD_{50} was identified.

- 4 (44%) of nine substances in the $2000 < LD_{50} \leq 5000$ mg/kg GHS toxicity category were correctly predicted; however, since a total of 11 substances were predicted for this category, the predictivity was 36% (4/11).
- 0 (0%) of 12 substances with $LD_{50} > 5000$ mg/kg were correctly predicted

The overall accuracy of the NHK NRU test method with the RC rat-only weight regression was 30% [14 /47]) (**Table 6-5**). Toxicity was overpredicted for 47% (22) and underpredicted for 23% (11) of the 47 substances, compared with *in vivo* toxicity categories (i.e., the GHS categories for the reference LD_{50} values in **Table 4-2**). For this analysis, in terms of each GHS toxicity classification category:

- 0 (0%) of four substances with $LD_{50} < 5$ mg/kg were correctly predicted
- 1 (14%) of seven substances in the $5 < LD_{50} \leq 50$ mg/kg GHS toxicity category was correctly predicted
- 3 (60%) of five substances in the $50 < LD_{50} \leq 300$ mg/kg GHS toxicity category were correctly predicted; however, since six other substances were also predicted for this category, the predictivity was 33% (3/9)
- 8 (89%) of nine substances in the $300 < LD_{50} \leq 2000$ mg/kg GHS toxicity category were predicted correctly; however, since 18 other substances were also predicted for this category, the predictivity was 31% (8/26)
- 1 (11%) of nine substances in the $2000 < LD_{50} \leq 5000$ mg/kg GHS toxicity category was correctly predicted
- 1 (8%) of 13 substances with $LD_{50} > 5000$ mg/kg was correctly predicted

Discordant Substances for Prediction of Toxicity Category by the 3T3 and NHK NRU Test Methods and the RC Rat-Only Weight Regression

Appendix L-2 shows the discordant substances for which the *in vitro* predicted GHS toxicity category did not match that based on the reference rodent LD_{50} data using the RC rat-only weight regression. The two *in vitro* NRU cytotoxicity test methods over- and under-predicted the GHS toxicity category for a similar number of substances, compared with the GHS toxicity categories for the reference LD_{50} values in **Table 4-2**. For the 3T3 NRU test method, the GHS toxicity category of 19 (63%) of 30 discordant substances was

overpredicted and the GHS toxicity category of 11 (37%) substances was underpredicted. For the NHK NRU test method, the GHS toxicity category of 22 (67%) of 33 discordant substances was overpredicted and the toxicity of 11 (33%) discordant substances was underpredicted.

6.3.3 Prediction of Toxicity Category by the 3T3 and NHK NRU Test Methods with the RC Rat-Only Weight Regression Excluding Substances with Specific Mechanisms of Toxicity

Table 6-6 shows the concordance of the observed and predicted GHS acute oral toxicity categories for each *in vitro* NRU test method using the geometric mean IC_{50} values (of the three laboratories) and the RC rat-only weight regression after excluding substances with specific mechanisms of toxicity (see **Table 6-3**). The formula for this regression was log LD_{50} (mg/kg) = log IC_{50} (μg/mL) x 0.357 + 2.194. Accuracy is the agreement of the *in vitro* predicted GHS toxicity categories with those based on the reference rat oral LD_{50} values from **Table 4-2**.

In Vitro – In Vivo Concordance for the 3T3 and NHK NRU Test Methods with the RC Rat-Only Weight Regression Excluding Substances with Specific Mechanisms of Toxicity

- The overall accuracy of the 3T3 NRU test method with the RC rat-only weight regression after excluding substances with specific mechanisms of toxicity was 46% (21/46 substances) (**Table 6-6**), compared to 35% (16/46 substances) when the complete RC rat-only weight regression was used (**Section 6.3.2** and **Table 6-5**). **Table 6-6** shows that GHS toxicity category was overpredicted for 24% (19) and underpredicted for 30% (11) of the 46 substances compared with the *in vivo* GHS toxicity categories for the reference LD_{50} values in **Table 4-2**.

In terms of each GHS toxicity classification category:

- 0 (0%) of the four substances with $LD_{50} < 5$ mg/kg were correctly predicted
- 1 (14%) of seven substances in the $5 < LD_{50} \leq 50$ mg/kg GHS toxicity category was correctly predicted

Table 6-6 Prediction of GHS Toxicity Categories[1] by RC Rat-Only Weight Regression Excluding Substances with Specific Mechanisms of Toxicity[2]

Reference Rodent LD_{50}[3]	3T3 NRU-Predicted Toxicity Category						Total	Accuracy	Toxicity Overpredicted	Toxicity Underpredicted
	< 5	5 – 50	50 – 300	300-2000	2000-5000	> 5000				
< 5	0	0	2	2	0	0	4[4]	0%	100%	0%
5 – 50	0	1	4	2	0	0	7[5]	14%	86%	0%
50 – 300	0	0	4	1	0	0	5[6]	80%	20%	0%
300 – 2000	0	1	1	7	0	0	9[7]	78%	0%	22%
2000 – 5000	0	0	0	3	6	0	9[8]	67%	0%	33%
> 5000	0	0	0	5	4	3	12[9,10]	25%	0%	75%
Total	0	2	11	20	10	3	46	46%	24%	30%
Predictivity	0%	50%	36%	35%	60%	100%				
Category Underpredicted	0%	50%	9%	40%	40%	0%				
Category Overpredicted	0%	0%	55%	25%	0%	0%				
Reference Rodent LD_{50}[3]	**NHK NRU-Predicted Toxicity Category**						**Total**	**Accuracy**	**Toxicity Overpredicted**	**Toxicity Underpredicted**
	< 5	**5 – 50**	**50 – 300**	**300 – 2000**	**2000 – 5000**	**> 5000**				
< 5	0	0	2	2	0	0	4[4]	0%	100%	0%
5 – 50	0	1	4	2	0	0	7[5]	14%	86%	0%
50 – 300	0	1	3	1	0	0	5[6]	60%	20%	20%
300 – 2000	0	1	0	8	0	0	9[7]	89%	0%	11%
2000 – 5000	0	0	0	5	4	0	9[8]	44%	0%	56%
> 5000	0	0	0	4	7	2	13[10]	15%	0%	85%
Total	0	3	9	22	11	2	47	38%	23%	38%
Predictivity	0%	33%	33%	36%	36%	100%				
Category Underpredicted	0%	67%	0%	41%	64%	0%				
Category Overpredicted	0%	33%	67%	23%	0%	0%				

[1]Globally Harmonized System of Classification and Labelling of Chemicals with LD_{50} in mg/kg (UN 2005).

569 < 5: $LD_{50} \leq 5$ mg/kg
570 5 – 50: $5 < LD_{50} \leq 50$ mg/kg
571 50 – 300: $50 < LD_{50} \leq 300$ mg/kg
572 300 – 2000: $300 < LD_{50} \leq 2000$ mg/kg
573 2000 – 5000: $2000 < LD_{50} \leq 5000$ mg/kg
574 > 5000: $LD_{50} > 5000$ mg/kg
575 [2]The RC rat-only weight regression excluding substances with specific mechanisms of toxicity is log LD_{50} (mg/kg) = log IC_{50}
576 (μg/mL) X 0.357 + 2.194.
577 [3]Reference rodent LD_{50} values from **Table 4-2**.
578 [4]Epinephrine bitartrate excluded because no rat LD_{50} was identified. Disulfoton and physostigmine excluded based on
579 mechanism of toxicity (see **Table 6-3**).
580 [5]Colchine excluded because no rat LD_{50} was identified. Endosulfan, parathion, potassium cyanide, and strychnine excluded based
581 on mechanism of toxicity (see **Table 6-3**).
582 [6]Dichlorvos, fenpropathrin, lindane, paraquat, phenobarbital, nicotine, and verapamil HCl excluded based on mechanism of
583 toxicity (see **Table 6-3**).
584 [7]Amitriptyline, atropine sulfate, caffeine, chloral hydrate, glutethimide, haloperidol, and procainamide HCl excluded based on
585 mechanism of toxicity (see **Table 6-3**).
586 [8]Carbon tetrachloride excluded because no laboratory attained sufficient toxicity for the calculation of an IC_{50}. Carbamazepine
587 excluded based on mechanism of toxicity (see **Table 6-3**).
588 [9]Methanol excluded because no laboratory attained sufficient toxicity for the calculation of an IC_{50}.
589 [10]Propylparaben excluded because no rat LD_{50} was identified.
590

- 4 (80%) of five substances in the $50 < LD_{50} \leq 300$ mg/kg GHS toxicity category were correctly predicted. Since seven other substances were also predicted for this category, predictivity was 36% (4/11).
- 7 (78%) of nine substances in the $300 < LD_{50} \leq 2000$ mg/kg GHS toxicity category were predicted correctly. Since a total of 20 substances were predicted for this category, the predictivity was 35% (7/20).
- 6 (67%) of nine substances in the $2000 < LD_{50} \leq 5000$ mg/kg GHS toxicity category were correctly predicted; the predictivity of this category was 60% (6/10)
- 3 (25%) of 12 substances with $LD_{50} > 5000$ mg/kg were correctly predicted. Since no other substances were predicted for this category, the predictivity was 100% (3/3).

Table 6-6 shows that the accuracy of the NHK NRU test method with the RC rat-only weight regression excluding substances with specific mechanisms of toxicity was 38% (18/47), compared to the 30% (14/47) accuracy when the complete RC rat-only weight regression was used (see **Table 6-5**). Toxicity was overpredicted for 23% (11) and underpredicted for 38% (19) of the 47 substances compared with the *in vivo* GHS categories for the reference LD_{50} values in **Table 4-2**. In terms of each GHS toxicity classification category:

- 0 (0%) of the four substances with $LD_{50} < 5$ mg/kg were correctly predicted
- 1 (14%) of seven substances in the $5 < LD_{50} \leq 50$ mg/kg GHS toxicity category was correctly predicted
- 3 (60%) of five substances in the $50 < LD_{50} \leq 300$ mg/kg GHS toxicity category were correctly predicted. Since six other substances were also predicted for this category, predictivity was 33% (3/9).
- 8 (89%) of nine substances in the $300 < LD_{50} \leq 2000$ mg/kg GHS toxicity category were predicted correctly. Since 14 other substances that did not match this category were also predicted, predictivity was 36% (8/22).
- 4 (44%) of nine substances in the $2000 < LD_{50} \leq 5000$ mg/kg GHS toxicity category were correctly predicted; the predictivity of this category was 36% (4/11)

- 2 (15%) of 13 substances with LD_{50} > 5000 mg/kg were correctly predicted. Since no other substances were predicted for this category, the predictivity was 100% (2/2).

Discordant Substances for the Prediction of Toxicity Category by the 3T3 and NHK NRU Test Methods and the RC Rat-Only Weight Regression Excluding Substances with Specific Mechanisms of Toxicity

Appendix L-2 shows the discordant substances for which the *in vitro* NRU predicted toxicity category did not match that based on the reference rodent LD_{50} data. The NHK NRU test method had four more discordant substances than the corresponding assay using 3T3 cells when the IC_{50} results were applied to the RC rat-only weight regression excluding substances with specific mechanisms of toxicity. For the 3T3 NRU test method, the GHS toxicity category of 19 (63%) of 30 discordant substances was overpredicted while the toxicity of 11 (37%) of 30 discordant substances was underpredicted compared with the *in vivo* GHS toxicity categories for the reference LD_{50} values in **Table 4-2**. For the NHK NRU test method, the toxicity of 22 (65%) of 34 discordant substances was overpredicted while the toxicity of 12 (35%) of 34 discordant substances was underpredicted.

6.3.4 Summary of the Regressions Evaluated

Table 6-7 summarizes the regressions evaluated in **Section 6.3** for accuracy in predicting the GHS acute oral toxicity categories (UN 2005) and the proportion of discordant substances for *in vitro* predictions of GHS toxicity categories. Accuracy for the 3T3 NRU test method was slightly lower than that for the NHK NRU test method for the RC millimole regression (i.e., 26% vs. 28%). Accuracy for the 3T3 NRU test method was higher than that for the NHK NRU test method for the RC rat-only weight regression (i.e., 35% vs. 30%) and the RC rat-only weight regression excluding substances with specific mechanisms of toxicity (i.e., 46% vs. 38%). The proportion of discordant substances for the 3T3 NRU test method was higher for the RC millimole regression (74%) than it was for the RC rat-only weight (65%) regression and the RC rat-only weight regression excluding substances with specific mechanisms of toxicity (65%). The proportion of discordant substances for the NHK NRU test method was similar for each regression (i.e., 70-72%). **Table 6-7** shows that the

difference between the proportions of discordant substances for the 3T3 and NHK NRU test methods widened with each subsequent regression (74% vs. 72% for the RC millimole regression, 65% vs. 70% for the RC rat-only weight regression, and 65% vs. 72% for the RC rat-only weight regression excluding substances with specific mechanisms of toxicity).

Table 6-7 Comparison of Regressions and *In Vitro* NRU Test Methods for Performance in Predicting GHS[a] Toxicity Categories

Regression	N[b]	Adjusted R^2	Accuracy	Discordant Substances[c]
RC –millimole units	347	0.450[d]	3T3 – 26% NHK – 28%	3T3 – 34/46 (74%) NHK – 34/47 (72%)
RC rat only –weight units[e]	282	0.322	3T3 – 35% NHK – 30%	3T3 – 30/46 (65%) NHK – 33/47 (70%)
RC rat only excluding substances with specific mechanisms of action – weight units[e]	232	0.353	3T3 – 46% NHK – 38%	3T3 – 30/46 (65%) NHK – 34/47 (72%)

[a]Globally Harmonized System of Classification and Labelling of Chemicals with LD_{50} in mg/kg (UN 2005).
[b]Number of substances used in regression.
[c]Proportion of substances evaluated.
[d]Calculated from RC data (i.e., regression not reported by Halle [1998]).
[e]From **Table 6-1**.

The highest accuracy for both *in vitro* NRU cytotoxicity test methods was attained when using the RC rat-only weight regression excluding substances with specific mechanisms of toxicity. The accuracy for the 3T3 NRU test method was 46%, which was greater than the accuracy of the 3T3 NRU with the RC millimole regression (26%) and with the RC rat-only weight regression (35%). The accuracy for the NHK NRU test method was 38% for the RC rat-only weight regression excluding substances with specific mechanisms of toxicity, 28% with the RC millimole regression, and 30% with the RC rat-only weight regression.

6.4 Strengths and Limitations of the *In Vitro* NRU Cytotoxicity Test Methods for *In Vivo* Toxicity Prediction

For each regression evaluated, the NRU basal cytotoxicity test methods tended to underpredict the toxicity of the most toxic substances and to overpredict the toxicity of the least toxic substances. The 3T3 and NHK NRU test methods were better at predicting the toxicity of substances with $50 < LD_{50} \leq 300$ mg/kg and $300 < LD_{50} \leq 2000$ mg/kg than

predicting the toxicity of substances with higher or lower LD_{50} values. The accuracy for the RC millimole regression and the RC rat-only weight regression for these toxicity categories was 67-100% for the 3T3 NRU and 33-83% for the NHK NRU data. Substances with higher or lower LD_{50} values were infrequently predicted correctly. The accuracy for substances with $LD_{50} \leq 50$ mg/kg (GHS toxicity categories for $LD_{50} \leq 5$ mg/kg and $5 < LD_{50} \leq 50$ mg/kg) was 0-17% for the 3T3 NRU and 0-50% for the NHK NRU with the same regressions. Accuracy for substances in the $2000 < LD_{50} \leq 5000$ and $LD_{50} > 5000$ mg/kg toxicity categories was 0-44% for the 3T3 NRU and 0-11% for the NHK NRU.

The RC rat-only weight regression calculated after removal of substances with specific mechanisms of toxicity improved the accuracy of GHS toxicity category predictions for substances with $LD_{50} > 2000$ mg/kg compared with the accuracy for the other regressions. The accuracy for substances in these categories was 25-67% for the 3T3 NRU and 15-44% for the NHK NRU. The RC rat-only weight regression excluding substances with specific mechanisms of toxicity did not increase the accuracy for substances with $LD_{50} < 2000$ mg/kg. However, the accuracy for substances in the $50 < LD_{50} \leq 300$ mg/kg and $300 < LD_{50} \leq 2000$ mg/kg categories using the RC millimole regression and the RC rat-only weight regression was already quite high. The accuracy for predicting these categories using the RC rat-only weight regression excluding substances with specific mechanisms of toxicity was 78-80% for the 3T3 NRU and 60-89% for the NHK NRU. The accuracy for predicting the toxicity categories for $LD_{50} \leq 5$ mg/kg and $5 < LD_{50} \leq 50$ mg/kg was 0-14% for both the 3T3 NRU test methods when using the RC rat-only weight regression excluding substances with specific mechanisms of toxicity.

The analysis of the 30 (3T3 NRU) to 31 (NHK NRU) discordant substances for the RC millimole regression to determine the physical, chemical, and biological characteristics associated with the discordant substances is presented in **Appendix L-1**. The analysis showed that 3 of 3 (100%) organophosphates were discordant in both test methods (10% of the 30 [3T3 NRU)] to 31 [NHK NRU] discordant substances). Other characteristics that seemed promising for characterizing RC millimole regression outliers were boiling point, molecular weight, and log K_{ow}. For boiling points > 200°C, 9/13 substances (69%) were

713 outliers for both the 3T3 results NHK NRU results (29 and 26% of the outliers, respectively).
714 The toxicity of seven of the nine (78%) outliers with boiling points > 200°C was
715 underpredicted by the RC millimole regression and the toxicity of the other two (22%)
716 substances was overpredicted. For molecular weight > 400 g/mole, 5/7 (71%) substances
717 were outliers using the 3T3 data and 3/7 (43%) were outliers using the NHK data (17 and
718 10% of the outliers, respectively). The toxicity of all the outliers with molecular weight >
719 400 g/mole was underpredicted by the RC millimole regression (5/5 [100%] for the 3T3
720 NRU and 3/3 [100%] for the NHK NRU). For log K_{ow} > 3, 9/12 (75%) substances were
721 outliers using the 3T3 data (30% of the outliers) and 8/12 (67%) substances were outliers
722 using the NHK data (26% of the outliers). The toxicity of 7/9 (78%) outliers (with
723 log K_{ow} > 3) for the 3T3 NRU assay was underpredicted by the RC millimole regression and
724 the toxicity of 2/9 (22%) outliers was overpredicted. The toxicity of 6/8 (75%) outliers (with
725 log K_{ow} > 3) for the NHK NRU assay was underpredicted by the RC millimole regression
726 and the toxicity of 2/8 (25%) outliers was overpredicted. Of the 21 substances with specific
727 mechanisms of toxicity that were not expected to be active in the 3T3 and NHK cell cultures,
728 13 (62%) were outliers. These substances represented 13/30 (43%) of the discordant
729 substances for the 3T3 NRU and 13/31 (42%) for the NHK NRU.
730
731 The lack of fit of individual substances to the regressions was not consistently related to their
732 insolubility in the 3T3 or NHK medium. Of the 25 substances that exhibited precipitates in
733 the 3T3 NRU assay, 11 (44%) substances were discordant (see **Table 5-8** for substances that
734 had precipitates and **Appendix L-1** for the analysis of discordant substances). The toxicity
735 of nine of the 11 (82%) substances was underpredicted by the RC millimole regression and
736 the toxicity of two of the 11 (18%) substances was overpredicted by the RC millimole
737 regression. Of the 24 substances that exhibited precipitates in the NHK NRU assay, 11
738 (46%) substances were outliers. The toxicity of nine of the 11 (82%) substances was
739 underpredicted by the RC millimole regression and the toxicity of two of the 11 (18%)
740 substances was overpredicted by the RC millimole regression.
741
742 Additionally, the lack of fit of individual substances to the RC millimole regression was not
743 consistently related to the fact that the test method systems had little to no metabolic

744 capability. Such a system would be expected to underestimate the toxicity of substances with
745 active metabolites. However, the toxicity of substances known to produce active metabolites
746 *in vivo* (listed in **Table 3-7**) was not necessarily underpredicted by the NRU assays. Of the
747 19 substances known to produce active metabolites *in vivo*, ten were discordant in the 3T3
748 NRU test method. Of these ten discordant substances, the toxicity of six (60%) was
749 underpredicted while the toxicity of four (40%) was overpredicted by the 3T3 NRU test
750 method. These ten discordant substances accounted for 33% of the 30 discordant substances
751 identified for the 3T3 NRU test method. Nine of the 19 substances known to produce active
752 metabolites *in vivo* were discordant for the NHK NRU test method. Of these nine discordant
753 substances, the NHK NRU assay underpredicted the toxicity of five (56%) substances and
754 overpredicted the toxicity of four (44%) substances. These nine discordant substances
755 accounted for 29% of the 31 discordant substances identified for the NHK NRU test method.
756
757 Similarly, Halle (1998) noted that the RC substances that required metabolic activation to
758 produce *in vivo* toxicity were not necessarily discordant substances (with respect to fit to the
759 RC millimole regression). Halle (1998) found that eight (50%) of the 16 substances that
760 required metabolic activation to product toxicity were discordant substances while eight
761 (50%) were not discordant (see **Table L3-3** in **Appendix L3**).
762
763 Some substances with low toxicity and low solubility could not be tested in the *in vitro* NRU
764 cytotoxicity assays because the amount of dissolved substance was inadequate to obtain an
765 IC_{50} value. In the 3T3 NRU test method, none of the laboratories obtained adequate toxicity
766 in any experiment with carbon tetrachloride and methanol. At least one laboratory failed to
767 achieve adequate toxicity with gibberellic acid and xylene. In the corresponding NHK assay,
768 no laboratory achieved adequate toxicity in any experiment with carbon tetrachloride, and at
769 least one laboratory could not achieve adequate toxicity with methanol, 1,1,1-trichloroethane,
770 and xylene.
771
772 Although the accuracy of the 3T3 and NHK NRU test methods for predicting *in vivo* toxicity
773 category was rather low when used with the RC millimole regression and the RC rat-only
774 weight regression, it was improved by removing substances with specific mechanisms of

toxicity that were not expected to be active in the 3T3 and NHK cell cultures. The evaluation of these *in vitro* NRU cytotoxicity test methods for predicting starting doses for acute systemic toxicity testing, thereby reducing and refining animal use, is provided in **Section 10**.

6.5 Salient Issues of Data Interpretation

One of the most important considerations for the 3T3 and NHK NRU test methods is getting good dose-response results. In addition to technical difficulties with these methods, such as occasional poor cell growth and the formation of NRU crystals, this validation study yielded observations of unusual dose-responses for certain substances.

The experimenter must be aware of dose-response anomalies and their causes in order to determine whether the dose-response can be better defined. For example, for substances such as aminopterin, which generally produced a biphasic dose-response using the log-dose spacing of the range-finder test, the experimenter must focus on the lowest concentration at which the substance produced 50% toxicity in order to perform the definitive testing with more closely spaced concentrations. In the definitive tests of such substances, the toxic response may plateau before producing 100% toxicity (i.e., 0% viability). The method used for the calculation of the IC_{50} must reflect an IC_{50} that is 50% inhibition of the control values rather than the midpoint of the highest and lowest response (as provided by the standard Hill function analysis).

Some substances, because of their low toxicity and/or low solubility, do not provide sufficient toxicity for the calculation of an IC_{50} value. Carbon tetrachloride, methanol, xylene, gibberellic acid, lithium carbonate and 1,1,1-trichloroethane failed to yield acceptable IC_{50} results in at least one laboratory due to insufficient toxicity/insolubility. All of these substances, with the exception of methanol, were reported to produce precipitate in the cell culture medium.

6.6 Comparison to Established Performance Standards

The *Guidance Document* method of evaluating basal cytotoxicity assays for use in predicting starting doses for acute oral toxicity assays provides the existing performance standard (ICCVAM 2001b) for the 3T3 and NHK NRU cytotoxicity test methods. The *Guidance Document* recommends testing10 to 20 reference substances from the RC in a candidate *in vitro* basal cytotoxicity assay to be used for predicting starting doses (ICCVAM 2001b). The substances should cover a wide range of toxicity and fit the RC prediction model (i.e., the linear regression line) as closely as possible. The IC_{50} results for the selected reference substances are used to calculate a new regression line with the LD_{50} values used by the RC. If the resulting regression is parallel to the RC millimole regression and within the ± log 5 (i.e., ± 0.699) prediction interval for the RC, the *Guidance Document* recommends using the cytotoxicity assay to predict starting doses for unknown substances to be tested in acute oral systemic toxicity assays.

One goal of the coded substance testing in Phases Ib and II of this study was to establish whether the results from the 3T3 and NHK NRU cytotoxicity test methods were consistent with the RC millimole regression. As discussed in **Section 3.4.1**, two of the major criteria for selecting the 12 coded substances tested in these phases from the 72 substances to be tested were (a) two substances must be included from each of the unclassified and classified GHS acute oral toxicity categories and (b) the substances must fit as closely to the RC millimole regression as possible. Unfortunately, the SMT could not identify 12 substances that fit both criteria since there was only one substance, aminopterin, in the $LD_{50} < 5$ mg/kg category that fit the RC millimole regression. The other substance chosen for testing for that toxicity category was sodium selenate. Since sodium selenate was not included in the RC, the SMT did not know how closely it would fit the RC millimole regression and it was not included in the regression analyses for Phases Ib and II.

The geometric mean log IC_{50} values from the 3T3 and NHK test methods from each laboratory were used with the oral log LD_{50} values from the RC (see **Appendices J-1 and J-3**) for the calculation of least squares linear regression analyses (see **Section 5.3**) for the

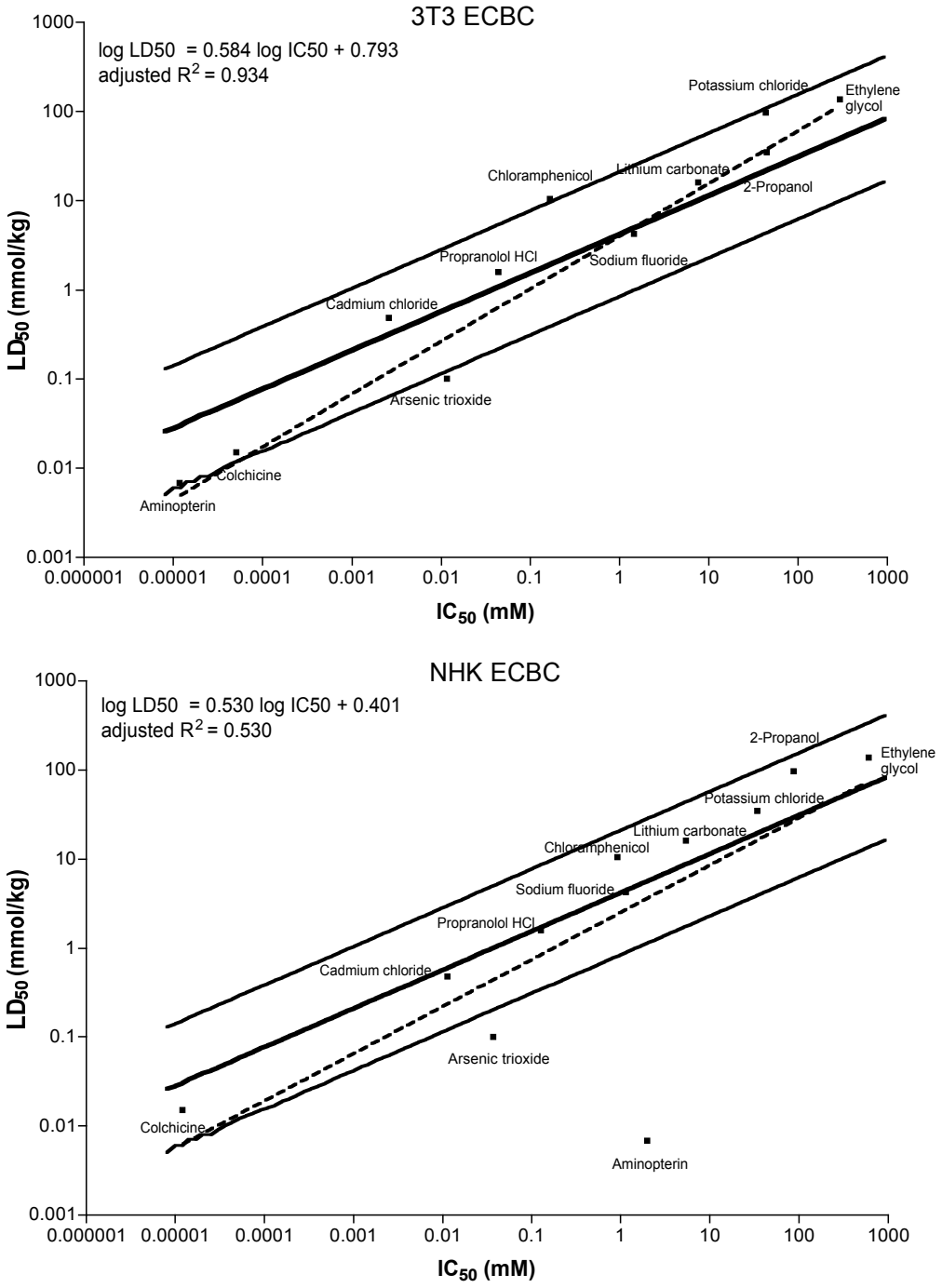
substances tested in Phases Ib and II. The slopes for all regressions were significantly different from zero with $p < 0.0001$. The adjusted R^2 values for the regressions from each laboratory, shown in **Table 6-8**, indicate that the 3T3 NRU test method produced better fitting regressions than the corresponding NHK assay (adjusted $R^2 = 0.934 – 0.947$ vs. 0.530 – 0.579). The relatively low adjusted R^2 values for the NHK assay were attributed to the much lower toxicity of aminopterin in that assay (see **Figures 6-6** to **6-8** and **Table 5-4**). The regressions were consistent with the RC millimole regression. **Table 6-8** shows that $p > 0.01$, the level of statistical significance, for all joint comparisons of slopes and intercepts with the RC millimole regression. The RC millimole regression slope and intercept were assumed to be constants for the comparison. A graphical comparison of the regressions with the RC millimole regression as suggested by the *Guidance Document* (ICCVAM 2001b) examples demonstrated that the regressions were generally within the RC millimole regression acceptance limits (see **Figures 6-6** to **6-8**). According to the *Guidance Document* (ICCVAM 2001b), basal cytotoxicity assays providing such consistency with the RC millimole regression are acceptable for predicting starting doses for *in vivo* acute oral systemic toxicity assays.

Table 6-8 Linear Regressions for Substances Tested in Phases Ib and II

	3T3 Millimole Regression			P-Values for Test Against RC Millimole Regression		
Laboratory	**Intercept**	**Slope**	**Adjusted R^2**	**Intercept**	**Slope**	**Joint[1]**
ECBC	0.793	0.584	0.934	0.202	0.014	0.040
FAL	0.709	0.598	0.947	0.497	0.008	0.024
IIVS	0.710	0.584	0.943	0.508	0.014	0.041
	NHK Millimole Regression			**P-Values for Test Against RC Millimole Regression**		
Laboratory	**Intercept**	**Slope**	**Adjusted R^2**	**Intercept**	**Slope**	**Joint[1]**
ECBC	0.401	0.530	0.530	0.484	0.547	0.620
FAL	0.429	0.548	0.579	0.519	0.450	0.569
IIVS	0.373	0.549	0.544	0.426	0.475	0.538

[1]Simultaneous comparison of slope and intercept. The RC slope and intercept were assumed to be constants.
ECBC – US Army Edgewood Chemical Biological Center; FAL – FRAME Alternatives Laboratory; IIVS – Institute for *In Vitro* Sciences

858 **Figure 6-6 *In Vitro – In Vivo* Regressions[1] for Phases Ib and II for ECBC**



859

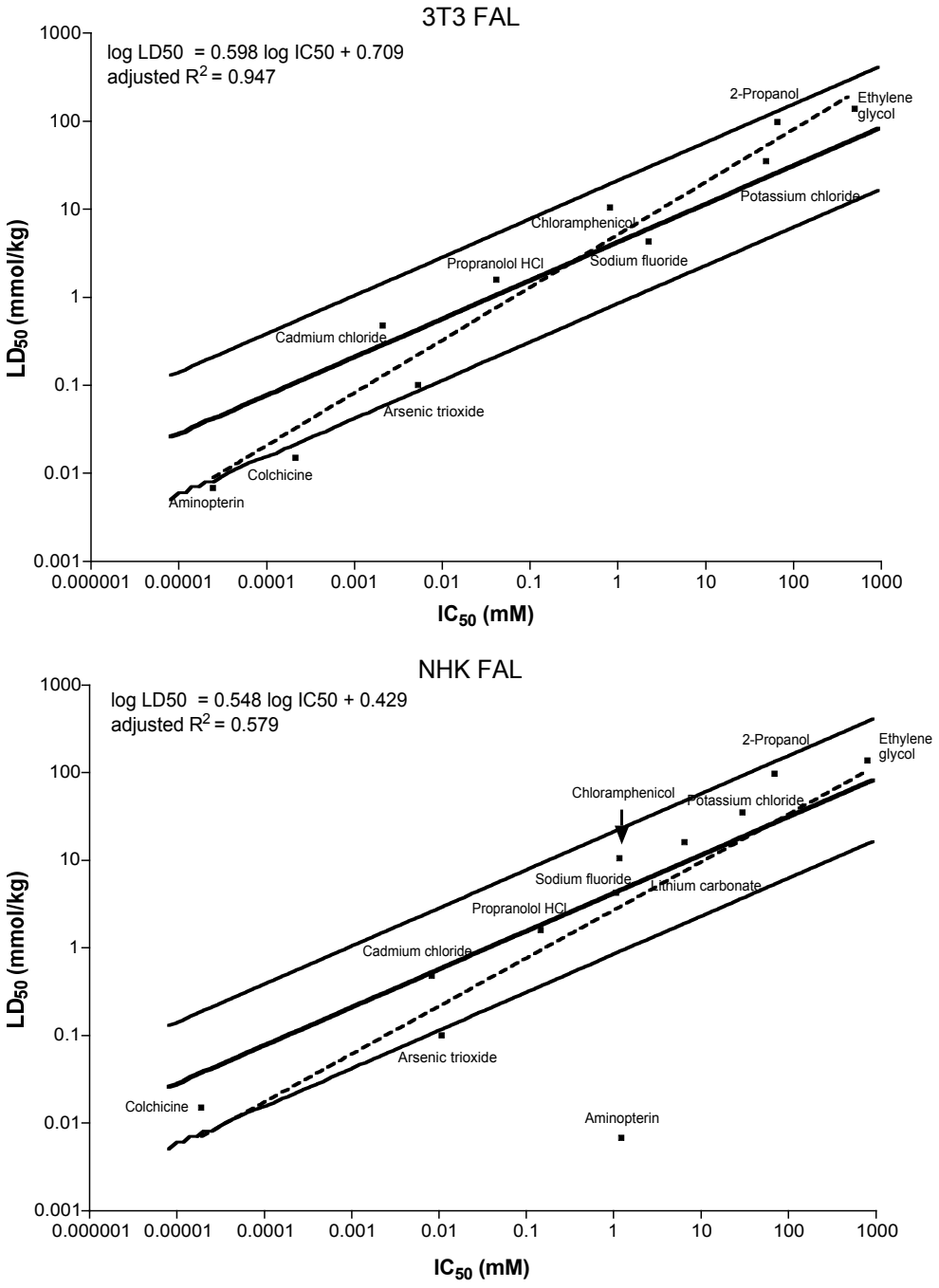
860 —— Solid Lines: RC millimole regression and acceptance limits - - - - - Dashed Line: Study Regression

861 [1]Regressions of substances tested in Phases Ib and II do not include sodium selenate since it was not included in
862 the RC.

863 ECBC: U.S. Army Edgewood Chemical Biological Command

864

865 **Figure 6-7 *In Vitro – In Vivo* Regressions[1] for Phases Ib and II for FAL**



866

867 —— Solid Lines: RC millimole regression and acceptance limits - - - - - Dashed Line: Study Regression

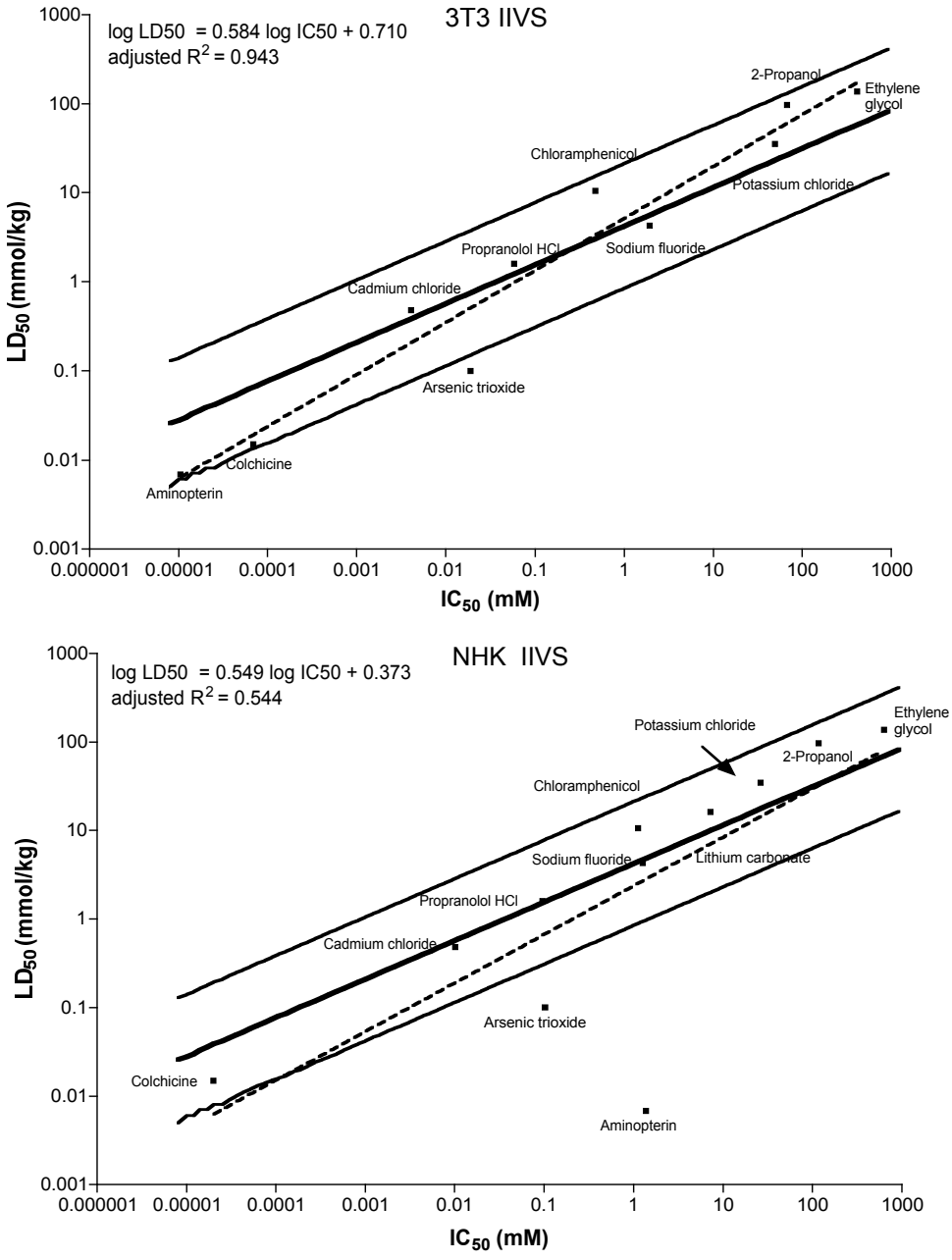
868 [1]Regressions of substances tested in Phases Ib and II do not include sodium selenate since it was not included in
869 the RC.

870 FAL: Fund for the Replacement of Animals in Medical Experiments Alternatives Laboratory

871

872

872 **Figure 6-8 *In Vitro – In Vivo* Regressions[1] for Phases Ib and II for IIVS**



873

874 —— Solid Lines: RC millimole regression and acceptance limits - - - - - Dashed Line: Study Regression
875 [1]Regressions of substances tested in Phases Ib and II do not include sodium selenate since it was not included in
876 the RC.
877 IIVS: Institute for In Vitro Sciences

878

879

880

6.7 Summary

The millimole regressions developed using the NICEATM/ECVAM IC_{50} and LD_{50} data were not significantly different from a regression for the 58 RC substances calculated using the RC data (F test; $p = 0.929$ for the 3T3 NRU regression and $p = 0.144$ for the NHK NRU regression). To improve the RC millimole regression with respect to the prediction of LD_{50} values by *in vitro* NRU IC_{50} values, regressions were developed using the RC data in weight units to exclude (1) mouse data (i.e., the RC rat-only weight regression) and (2) substances with mechanisms of toxicity that were not expected to be active in the 3T3 and NHK cell cultures (i.e., the RC rat-only regression excluding substances with specific mechanisms of toxicity regression).

Accuracy in predicting GHS acute toxicity category using these *in vitro* NRU cytotoxicity test methods was 26% (12/46) for the 3T3 NRU and 28% (13/47) for the NHK NRU with the RC millimole regression. Accuracy with the RC rat-only weight regression improved to 35% (16/46) for the 3T3 NRU and 30% (14/47) for the NHK NRU. Accuracy was higher for substances with $50 < LD_{50} \leq 2000$ mg/kg compared to substances with higher or lower toxicity. For these two regressions, the accuracy of predicting the $50 < LD_{50} \leq 300$ mg/kg and $300 < LD_{50} \leq 2000$ mg/kg categories for the 3T3 and NHK NRU was 67-100% and 50-100%, respectively. The accuracy of predicting the $LD_{50} \leq 5$ mg/kg and $5 < LD_{50} \leq 50$ mg/kg categories was 0-17% for the 3T3 NRU and 0-50% for the NHK NRU. The accuracy for substances with $2000 < LD_{50} \leq 5000$ mg/kg and $LD_{50} > 5000$ mg/kg was 0-67% and 0-44% for the 3T3 and NHK NRU data, respectively.

Examination of outliers for the RC millimole regression by chemical class showed that 3 of 3 (100%) organophosphates were outliers in both test methods. Other characteristics that seemed promising for characterizing RC outliers were boiling point, molecular weight, and log K_{ow}. For boiling points > 200°C, 9/13 (69%) substances were outliers for both the 3T3 and NHK NRU results. For molecular weight > 400 g/mole, 5/7 (71%) substances were

910 outliers using the 3T3 data and 3/7 (43%) were outliers using the NHK data. For log K_{ow} >
911 3, 9/12 (75%) substances were outliers using the 3T3 data and 8/12 (67%) substances were
912 outliers using the NHK data.
913
914 The lack of fit of individual substances to the RC millimole regression was not consistently
915 related to substance insolubility in the 3T3 or NHK medium or to the fact that the test
916 method systems had little to no metabolic capability. Of the substances that exhibited
917 precipitates, 11/25 (44%) substances were discordant with the 3T3 NRU assay and 11/24
918 (46%) were discordant with the NHK NRU assay. Also, although the 3T3 and NHK cells
919 have little to no metabolic capability, the toxicity of substances known to produce active
920 metabolites *in vivo* was not necessarily underpredicted by these assays. Of the 19 substances
921 known to produce active metabolites *in vivo*, ten (53%) were discordant in the 3T3 NRU test
922 method. Of these ten discordant substances, the toxicity of six (60%) was underpredicted
923 while the toxicity of four (40%) was overpredicted by the 3T3 NRU test method. These ten
924 discordant substances accounted for 33% of the 30 discordant substances identified for the
925 3T3 NRU test method. Similarly, nine (47%) of the 19 substances known to produce active
926 metabolites *in vivo* were discordant for the NHK NRU test method. Of these nine discordant
927 substances, the NHK NRU assay underpredicted the toxicity of five (56%) substances and
928 overpredicted the toxicity of four (44%) substances. These nine discordant substances
929 accounted for 29% of the 31 discordant substances identified for the NHK NRU test method.
930
931 The examination of outliers based on mechanism of toxicity lead to the development the RC
932 rat-only weight regression excluding substances with specific mechanisms of toxicity. Of the
933 21 substances with specific mechanisms of toxicity that were not expected to be active in the
934 3T3 and NHK cell cultures, 13 (62%) were outliers. These substances represented 13/30
935 (43%) of the discordant substances for the 3T3 NRU test method and 13/31 (42%) for the
936 NHK NRU test method. The RC rat-only weight regression excluding substances with
937 specific mechanisms of toxicity improved the accuracy from 26% (12/46) for the RC
938 millimole regression to 46% (21/46) for the 3T3 NRU test method and from 28% (13/47) to
939 38% (18/47) for the NHK NRU test method.
940

The RC rat-only weight regression calculated after removal of substances with specific mechanisms of toxicity improved the accuracy (compared with the RC millimole regression) for predicting most toxicity categories. It did not improve the accuracy of category prediction for substances with $LD_{50} < 50$ mg/kg or for substances with $300 < LD_{50} \leq 2000$ mg/kg. The following changes in accuracy for the various toxicity categories, compared with the RC millimole regression, occurred:

- $LD_{50} \leq 5$ mg/kg– 0% to 0% for both 3T3 and NHK NRU
- $5 < LD_{50} \leq 50$ mg/kg – 17% to 14% for the 3T3 NRU and 50% to 14% for the NHK NRU
- $300 < LD_{50} \leq 2000$ mg/kg – 100% to 78% for the 3T3 NRU and 100% to 89% for the NHK NRU
- $2000 < LD_{50} \leq 5000$ mg/kg – 0% to 67% for the 3T3 NRU and 9% to 44% for the NHK NRU
- $LD_{50} > 5000$ mg/kg – 10% to 25% for the 3T3 NRU and 0% to 15% for the NHK NRU

[This Page Intentionally Left Blank]